Statistical Methods
Identification of Differentially Expressed Genes and Class Prediction Using Mutual Information and Total Number of Misclassification Analyses

Amir Ben-Dor
Zohar Yakhini

amir_ben-dor@agilent.com
zohar_yakhini@agilent.com

Scope

In this report we describe an analysis performed to evaluate the relationship between DNA level information and expression level characteristics in hereditary breast cancer samples. We evaluate the performance of gene expression based genotyping methodologies. We provide lists of genes that are highly relevant to the expression level manifestation of the genetic variation. We also describe the statistical methods in some detail but refer the mathematically inclined readers to more complete related work on the methods.

Data

The data consists of 22 breast cancer samples. 7 are BRCA1 mutants, 8 are BRCA2 mutants, and 7 are sporadic. Expression levels are measured for 3226 genes, using cDNA microarrays and differentially labeled samples.

Highlights of the Results

- For both mutation types (BRCA1 and BRCA2) there is a statistically significant over-abundance of highly relevant genes in the dataset. We have compiled two lists of highly relevant genes, one for each mutation type.
- We show that *in-silico* gene-expression based genotyping success rates depend on the set of genes actually used to make decisions. We can correctly call up to 95% (21/22) of the samples for both mutation types, using sets of significant genes. The dependence on the

particular set being used is, however, not very strong. High success rates (such as 20/22) are obtained for a range of selected subsets.

- We identify one sporadic sample that manifests a strong BRCA1 phenotype in its mRNA expression profile. This sample might be a good candidate for further investigation.

Methods Outline

In the Analysis, we treated the two loci (BRCA1 and BRCA2) independently. That is, we first considered the genotype information for the BRCA1 locus (mutation vs. the wild type), ignoring the information regarding the BRCA2 locus. Similarly, when we analyzed the genotype information for the BRCA2 locus, we ignored the BRCA1 locus. This approach allows for samples to have none of the mutations, one of them or both.

Our approach consists of two independent parts (details below) - ***Gene scoring***, aimed at selecting genes with statistically significant relevance to the genotype at the locus under consideration; and ***in-silico genotyping assays***, aimed at simulating the prediction of the genotype of an unknown sample based on its mRNA expression profile. Gene scoring also provides a rigorous statistical evaluation of the overabundance of highly relevant genes, as explained below.

Methods

Scoring Genes for Relevance

Attaching a measure of relevance to each gene in classified expression data is useful in several ways. Seeking small sets of genes that can jointly serve as a classifier and as a basis for the development of diagnostic assays, one can choose amongst the more informative genes found in preliminary more comprehensive studies. Highly informative genes that are parts of known biochemical pathways give insight into the processes that underlay the differences between classes. Highly informative genes (or ESTs) of unknown function suggest new research directions.

We evaluated two different gene-scoring methods: *Information-score* (*Info* for short) *and Thresholded-number-of-Misclassifications* (*TNoM* in short). Full definitions of these scores are given in this section and a more comprehensive discussion of scoring methods can be found in Tech. Report AGL-2000-13, Agilent Labs, Agilent Technologies, 2000, http://www.labs.agilent.com/resources/techreports.html.

Roughly speaking, the *info* score of a gene **g** is the amount of uncertainty that is left regarding the genotype of an unknown sample **U**, after we learn the expression level of **g** in **U**. Thus, the lower the info score is, the more relevant the gene is for genotyping. In addition, we attach a significance level (p-value) to each *info*-score. The p-value is the probability to get this *Info* score (or better) at random, as described below in further detail (Similar analysis is performed for the *TNoM* scores, which corresponds to the number of “errors” - tissues that express differently than their class.)

Having sorted the genes according to their *Info* score, we report whether there is an over-abundance of informative genes in the date set. We plot the actual number of genes in the data set (as a function of their *Info* score), together with the number of genes with the same *Info* score expected in random data. The sorted list of genes also serves to identify genes that are most relevant to the studied phenomenon and thus to point at promising research directions.

Using *Info* or other scores in analyzing actual gene expression data one does encounter many genes that are strongly indicative of various classes of samples. It is important to statistically assess such findings in a sound manner. In analogy, when two highly homologous sequences are encountered it is more important to consider the associated p-value (such as returned by BLAST, for example) than the actual homology distance under consideration.

To evaluate the statistical significance of gene relevance we need to develop a null model. We then estimate the probability of a gene scoring better than some fixed level *s* in randomly labeled data (according to said model). This number is the *p-value* corresponding to the scoring method in effect and the given level *s*, under the prevailing null model.

Specific advantages and applications of statistically sound relevance scoring are:

- Genes with very low p-values are very rare in random data and their relevance to the studied phenomenon is therefore likely to have biological, mechanistic or protocol reasons. Genes with low p-values for which the latter two options can be ruled out are interesting subjects for further investigation and are expected to provide deeper insight into the studied phenomena.
- Assessment of putative subclasses in the data. p-Values for relevance scores allow for comparing a candidate partition of the sample set in the data to a uniformly drawn partition of the same composition, in terms of the abundance of very informative genes. This serves to underline the biological meaning of a partition. In other words, this comparison statistically validates a candidate partition as having properties that would only very rarely occur for random partitions. This analysis was instrumental in a melanoma gene expression study reported in Bittner et al, Nature, 2000. The authors applied relevance scores for the assessment of the statistical significance of a putative coetaneous melanoma subtype and for selecting differentially expressed genes.
- Assessment of the information content of classified gene expression data by considering the deviation of the numbers of relevant genes in the data from those

expected for random classes. This analysis is exemplified in the Results Section below.

- p-Values provide a common platform for comparing individual gene relevance across different datasets, different scores and different partitions of the same data.
- In actual gene expression data it is often the case that expression levels for some genes are not reported for some samples. This is typically due to technical measurement problems. The result is that the mixture of labels that needs to be considered is dependent on the gene in question. Obviously, a given *Info* or *TNoM* score has a different significance level for a 20:20 mixture then it does for a 20:5 mixture. When selecting a subset of genes as a classification platform or when looking for insight into the studied biological process we should therefore consider the relevance of each gene in the context of the appropriate mixture. Absolute score values do not provide a uniform figure of merit in this context. We use p-values as a uniform platform for such comparisons, as they do account for the mixture that defines the model.

In this current study of hereditary breast cancer expression profiles we employed efficient methods for calculating exact p-values for the *Info* and *TNoM* scores defined below. These exact calculations allow for the information overabundance analysis.

Scoring Methods: Info and TNoM

In this section we describe the two scoring used in this current study of hereditary breast cancer. To this end we need some notations. Let **k** denote the number of tissues, consisting of **a** tissues from class **A**, and **b** tissues of class **B**. Assume we want to score a gene **g** for relevance with respect to the **A:B** partition of the tissues. Intuitively, **g** is relevant to the tissue partition if it is either over-expressed in class **A** tissues (compared to class **B** tissues) or vice-versa.

To formalize the notion of relevance, we consider how **g** expression levels in class **A** tissues interlace with its expression levels in class **B** tissues. Denote by t_i the i-th tissue ranked according to the expression level of **g** (that

is, **g** express minimally in t_1 and maximally in t_k). We define the ***rank vector***, ***v***, of **g** to be a {-,+} vector of length **k**, as follows:

$$v_i = \begin{cases} + & if\ t_i \in \mathbf{A} \\ - & if\ t_i \in \mathbf{B} \end{cases}$$

For example, if **g**'s expression levels in class **A** are {10,20,30,50,60,70,110,140}, and **g**'s expression levels in class **B** are {40,80,90,100,120,130,150} then

$$v = \langle +,+,+,-,+,+,+,-,-,-,+,-,-,+,- \rangle \qquad (1)$$

Note that the rank vector *v* captures the essence of the differential expression profile of **g**. If **g** is under-expressed in class **A**, then the positive entries of *v* are concentrated in the left hand side of the vector, and the negative entries are concentrated at the right hand side. Similarly, for the opposite situation. Therefore, the relevance of **g** increases as the homogeneity within the left hand side of *v*, and the homogeneity within the right hand side of *v* increase.

Two natural ways to define the homogeneity on the two sides, and to combine them into one score, lead to the two scoring method, ***TNoM*** and ***Info***. In both cases the score of *v* corresponds to the maximal combined homogeneity over all possible ways to break v to two parts.

TNoM score

Define the ***Min-Cardinality***, of a {-,+} vector x, to be the cardinality of the minority symbol in x. That is

$$MC(x) = \min\{\#_-(x), \#_+(x)\}.$$

The ***TNoM*** score of a rank vector v is defined as

$$TNoM(v) = \min_{x,y=v}\{MC(x) + MC(y)\}$$

For example, for the rank vector v in (1), the best partition of v into two parts is $v = \langle +,+,+,-,+,+,+ \| -,-,-,+,-,-,+,- \rangle$, and thus, $TNoM(v) = 1 + 2 = 3$.

Info Score

Let x be a {-,+} vector, and let p denote the fraction of positive entries in x. The ***entropy*** of x, is defined as

$$Ent(x) = H(p) = -p\log(p) - (1-p)\log(1-p).$$

The Info score of a rank vector v is defined to be the minimal weighted sum of the entropy of its two parts. I.e.,

$$Info(v) = \min_{x,y=v}\left\{\frac{|x|}{|v|}Ent(x) + \frac{|y|}{|v|}Ent(y)\right\}.$$

Using the same rank vector v from Equation (1), the best partition with respect to the Info score happens to be the same as the one above. We get

$$Info(v) = \frac{7}{15}H(\frac{6}{7}) + \frac{8}{15}H(\frac{2}{8}) = 0.7088$$

Information Overabundance Analysis

Given an *Info* score level ***s*** and a classified expression dataset with ***a*** samples of one class and ***b*** of another, consider the p–value of *s*:

$$\text{p-Val}(s;a,b) = \text{Prob}(\ Info(\mathbf{V}) \le s\ ),$$

where ***V*** is uniformly drawn rank vector over all {-,+} vectors with ***a*** + symbols and ***b*** – symbols.

The expected number, E(*s*), of genes with $Info(g) \le s$ is given by p-Val(*s*;*a*,*b*)·*n*, where *n* is the total number of genes in the dataset.

Information overabundance analysis compares E(*s*) to the actual number of genes with $Info(g) \le s$ observed for the classification under consideration, A(*s*). Our confidence in the significance of a given classification increases monotonically with the deviation E(*s*) > A(*s*), for small values of *s*. Exact p-value calculations (as opposed to permutation test approximation) enable the calculation of E(*s*). A(*s*) is computed directly from the data.

The above discusses information overabundance analysis using *Info*. The same analysis also applies to *TNoM* and is employed in this study.

In-Silico Genotyping Assay

In-silico genotyping assay employs a set of significant genes to predict the genotype of an unknown sample. We evaluated the performance of several *in-silico* genotyping assays (*Clustering-based-Classification*, *Nearest-Neighbor*, and *Voting*). As all methods performed similarly, we report here the results only for the (slightly superior) *Voting* method. In this method, each of the selected genes, **g**, casts a vote regarding the genotype of the unknown sample, **U** (based on **g**'s expression level in **U**). The collection of votes is used to predict the genotype of **U**.

To assess the predictive power of this method, we have performed Leave-One-Out-Cross-Validation (LOOCV) simulations. In this validation procedure we hide the genotype of one sample, **U**, and use the classification method (i.e., select a set significant genes, and perform a voting among them) to predict the genotype of the sample **U**. The predicted genotype is then compared to the real genotype. We repeat this process for all samples and record the success rate.

Clearly, the performance of the genotyping assay depends on the set of significant genes employed (to participate in the voting). Therefore, we select different sets of significant genes by setting different p-value threshold values. For example, a threshold value of 0.01, corresponds to selecting all genes with significance level of 0.01 (or better). In the results section we plot the success rate of the LOOCV experiments as a function of the p-value threshold used.

Results

Score Distributions, Information Overabundance

Figure 1 and Figure 2 depict the differences between the expected number of genes with a given *Info* score (or better) and the actual number of such genes in the data. Informative genes are significantly more abundant in the data than what would be expected at random. For example, there are 69 genes with *Info* score ≤ 0.405, for BRCA1, while only about 9 are expected. Similarly, there are 6 genes there with *Info* score ≤ 0.2 while about 0.3 are expected. This calculation does not assume any degree of gene independence.

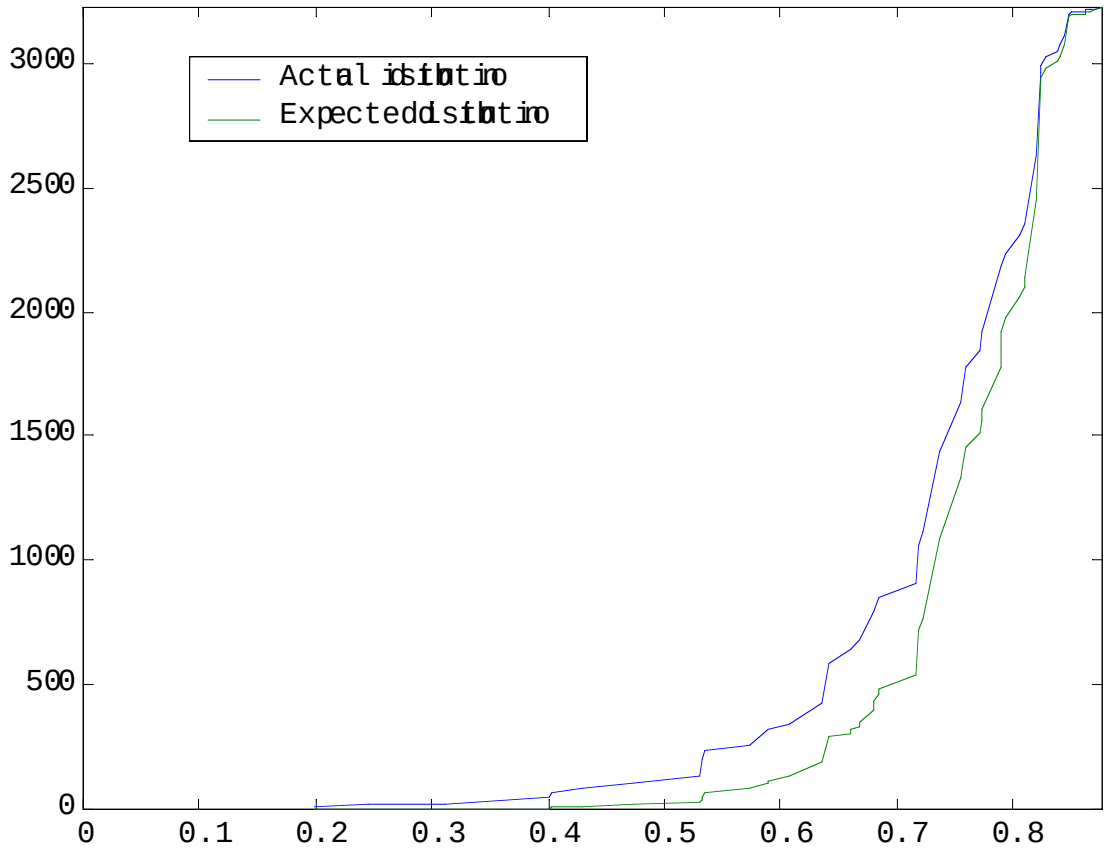


Figure 1 Distribution of genes *info* scores for BRCA1 locus

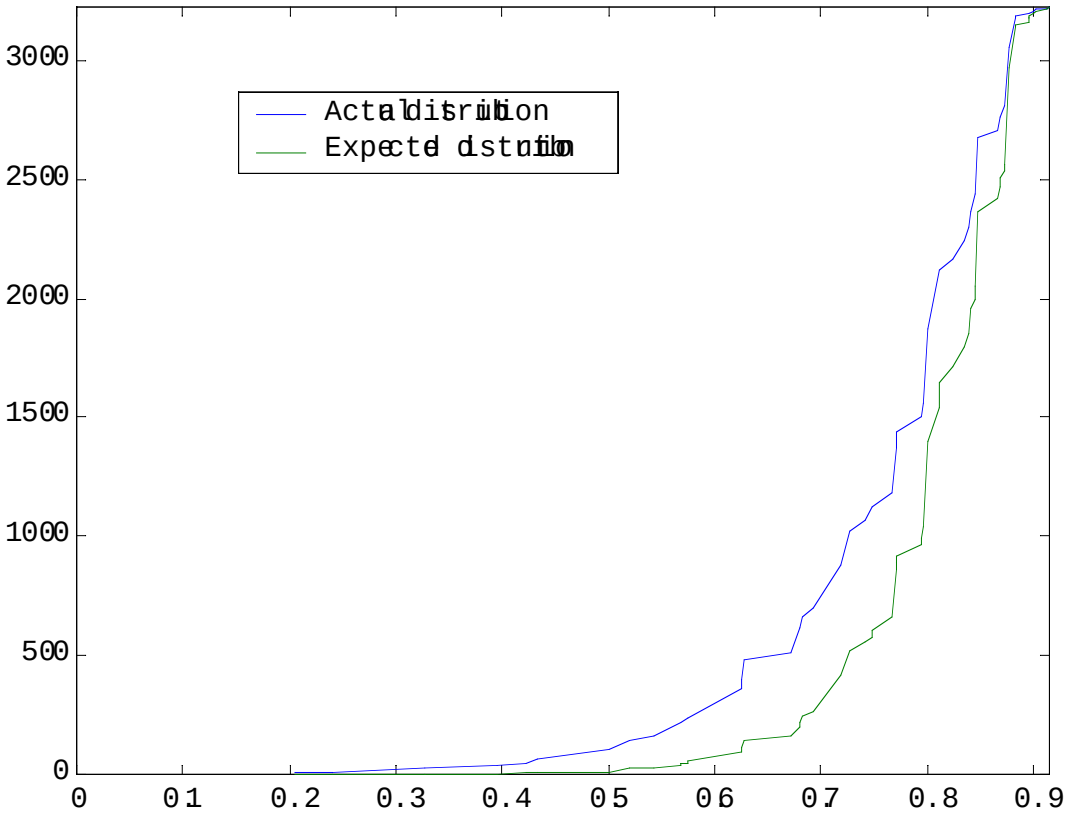


Figure 2 Distribution of genes info scores for BRCA2 locus

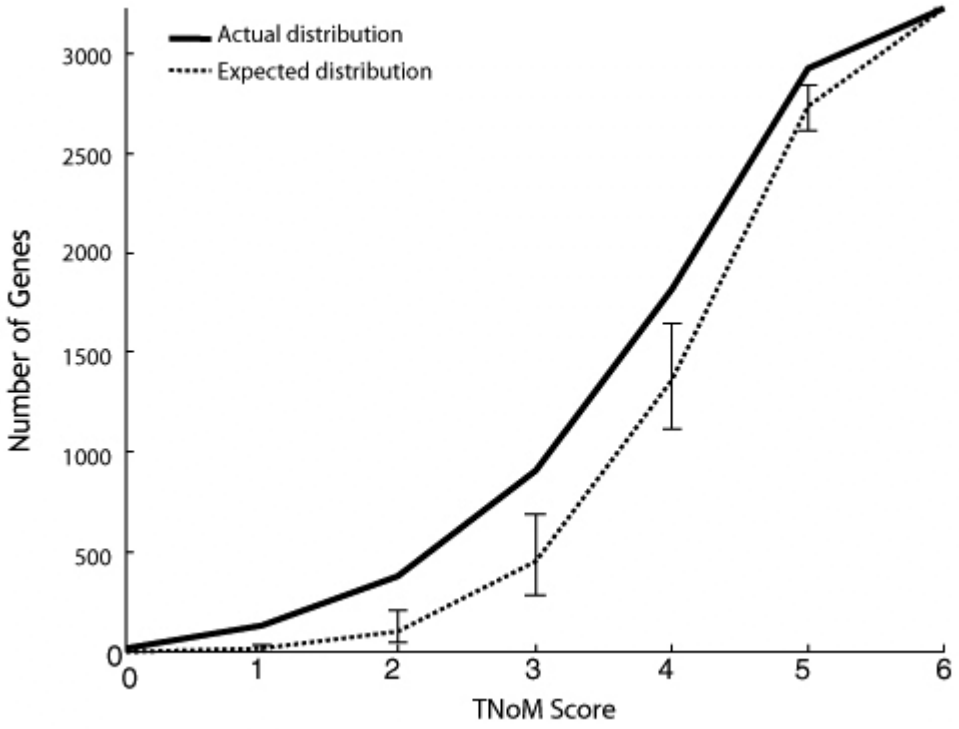


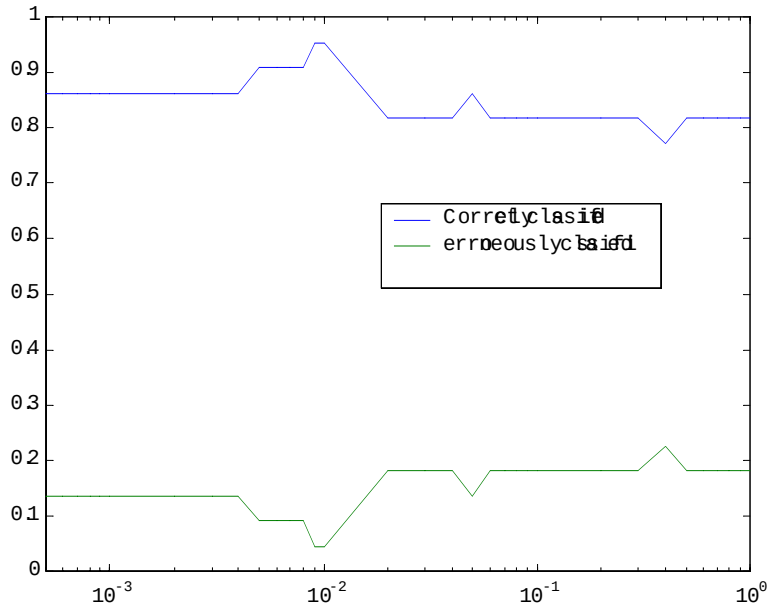
Figure 3 Information overabundance analysis BRCA1 vs. the rest of the sample, using the *TNoM* relevance score. Error bars correspond to 95% confidence intervals, derived from computer generated uniformly drawn partitions of samples. 1000 random partitions were drawn. The centers of the distributions are rigorously calculated as indicated above.

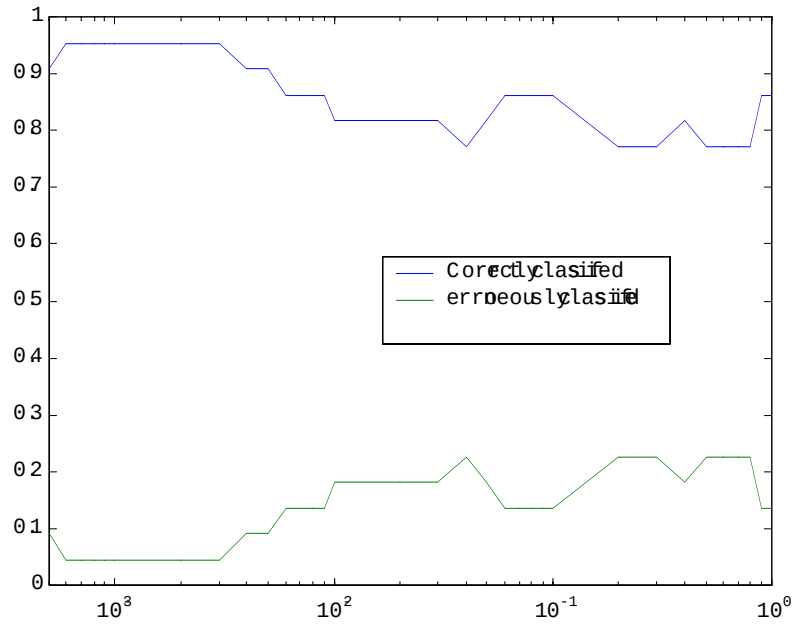
Lists of Relevant Genes

Please see adjacent Excel sheets for the relevance scores for all genes in the data.

LOOCV Results

Figures 3 and 4 below depict the results of LOOCV simulations for the BRCA1 and BRCA2 loci respectively. In both, we plot the fraction of correctly classified (and erroneously classified) samples as a function of the p-value threshold (in logarithmic scale).





Phenotype/Genotype Inconsistency

In our analysis, the expression profile of most tissues was consistent with the genotype of the sample (for different p-value thresholds). One notable exception, though, is the sample "Sporadic 1321". This sample exhibits a strong BRCA1 expression profile - all sets of significant genes called this sample as BRCA1.

References

- *Tissue classification with gene expression profiles* Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z, *JOURNAL OF COMPUTATIONAL BIOLOGY* 7: (3-4) 559-583 2000
- *Scoring genes for relevance, Ben-Dor A, Friedman N,* Yakhini Z., Tech. Report AGL-2000-13, Agilent Labs, Agilent Technologies, 2000, http://www.labs.agilent.com/resources/techreports.html.
- *Molecular classification of cutaneous malignant melanoma by gene expression profiling*, Bittner M, Meitzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V, Hayward N, Trent J *NATURE* 406: (6795) 536-540 AUG 3 2000